4. 로지스틱 회귀모형

2) 추론

2. 추론

- 로지스틱 회귀모형에서 사용할 수 있는 검정
  1. Wald Test
  2. Likelihood Ratio Test(LRT)
  3. Score Test

  - 세 검정 모두 likelihood function에 의한 방식
  - 자세한 이론적 배경은 생략
  - 근사적으로(표본크기가 무한대로 커지는 경우) 세 검정은 모두 동일한 결과를 보이는 것으로 알려져 있음
  - 실제 데이터를 대상으로 하는 경우에 약간 다른 결과 산출

1) Wald Test에 의한 개별회귀 계수 추론

1.1) 개별 회귀 계수 검정

- $H_0: \beta_j = 0,\ H_1: \beta_j \neq 0$
- Wald test 검정 통계량

$$Z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

- 귀무가설에서 근사적 정규분포
- SAS에서는 검정 통계량을 Z 대신 Z^2으로 사용. 이 경우 분포는 $\chi^2(1)$
- 점 추정량과 표본오차의 값은 모두 반복 계산으로 산출
- 경우에 따라 표본오차가 과도하게 크게 계산되기도 함

- 예제 3.1: 예제 2.1 자료에 대한 Wald test

```
> fit <- glm(lfp ~ ., family=binomial, data=Mroz)
> summary(fit)
```

표 3.1과 비교

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
k5          -1.462913   0.197001  -7.426 1.12e-13 ***
k618        -0.064571   0.068001  -0.950 0.342337
age         -0.062871   0.012783  -4.918 8.73e-07 ***
wcyes        0.807274   0.229980   3.510 0.000448 ***
hcyes        0.111734   0.206040   0.542 0.587618
lwg          0.604693   0.150818   4.009 6.09e-05 ***
inc         -0.034446   0.008208  -4.196 2.71e-05 ***
```

1.2) 개별 회귀계수의 신뢰구간 추정

$$\hat{\beta}_j \pm z_{1-\alpha/2} SE(\hat{\beta}_j)$$

- 예제 3.1:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
k5          -1.462913   0.197001  -7.426 1.12e-13 ***
k618        -0.064571   0.068001  -0.950 0.342337
age         -0.062871   0.012783  -4.918 8.73e-07 ***
wcyes        0.807274   0.229980   3.510 0.000448 ***
hcyes        0.111734   0.206040   0.542 0.587618
lwg          0.604693   0.150818   4.009 6.09e-05 ***
inc         -0.034446   0.008208  -4.196 2.71e-05 ***
```

- β_1에 대한 95% 신뢰구간 추정

```
> c(-1.4629-1.96*0.197, -1.4629+1.96*0.197)
[1] -1.849 -1.077
```

2) Deviance를 이용한 추론

- 모형의 적합도(Goodness of fit)
  - 주어진 모형에 의하여 추정된 반응변수의 적합값과 반응변수의 실제 관찰값과의 일치 정도를 의미
  - 적합도를 나타내는 측도는 다수 존재
  - Likelihood function : 이항반응변수의 로지스틱 회귀모형에 대한 적절한 적합도 측정 도구
    - 단독적으로는 의미를 부여하기 어려움
    - 현재 모형과 완전 모형(주어진 자료를 완전하게 설명하는 모형)의 likelihood function 값 비교로 현재 모형의 적합도를 측정

2-1) Deviance의 정의

- $D = -2\log(\hat{L}_c/\hat{L}_f) = -2\left[\log \hat{L}_c - \log \hat{L}_f\right]$

  $\hat{L}_c$: current(현재) 모형의 maximized likelihood

  $\hat{L}_f$: full(완전) 모형의 maximized likelihood

  - Deviance: 현재 모형과 완전 모형의 적합도 차이. 현재 모형에 의한 반응변수의 추정값과 실제 관찰값과의 일치 정도를 표현

  - 일반적으로 deviance D는 근사적인 χ^2 분포를 함

- 이항반응변수의 경우 Deviance

$$D = -2\sum\{\hat{\pi}_i \ \text{logit}(\hat{\pi}_i) + \log(1-\hat{\pi}_i)\}$$

  - 추정값만의 함수: 관찰값과의 비교 불가능
  - 현재 모형의 Deviance만으로는 적합도 표현 불가능
  - 분포: 일반적인 경우와 다르게 χ^2 분포 사용 불가능

2-2) 두 nested 모형의 deviance 비교

- nested 모형: B 모형에 몇 개의 항을 추가된 것이 A 모형일 때 두 모형의 관계

  - 예: 한 개 항 추가

    A 모형: $\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

    B 모형: $\text{logit}(\pi) = \beta_0 + \beta_1 X_1$

  - 예: 여러 항 추가

    A 모형: $\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p$

    B 모형: $\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$

- 두 nested 모형의 deviance 차이: 추가된 변수로 인한 적합도 향상 정도 측정

- 두 nested 모형의 deviance 차이

  - D_L: large 모형(A 모형)의 deviance
    D_S: small 모형(B 모형)의 deviance

  - 두 nested 모형의 deviance 차이: $D_S - D_L$

  - Large 모형에 추가된 변수가 적합도 향상에 유의적인 효과가 있는지를 검정하는 것은 중요한 문제

  - $D_S - D_L$의 분포: 근사적으로 χ^2분포를 하며 자유도는 두 모형의 모수 개수 차이(이항반응변수의 경우에도 적용)

2-3) Deviance에 의한 가설 검정

- 현재 모형: $\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- 검정 가능 가설:

1. 회귀모형 유의성($H_0: \beta_1 = \cdots = \beta_p = 0$)
2. 2개 이상 회귀계수 유의성(예; $H_0: \beta_1 = \beta_2 = 0$)
3. 개별 회귀계수 유의성(예; $H_0: \beta_1 = 0$)

- 예제 3.2 & 3.3:

1) $H_0: \beta_1 = \cdots = \beta_7 = 0$ (예제 3.2(3) & 3.3(1))

```
> fit <- glm(lfp ~ ., family=binomial, data=Mroz)
> fit_0 <- glm(lfp ~ 1, family=binomial, data=Mroz)
```

```
> anova(fit_0, fit, test="Chisq")
Analysis of Deviance Table

Model 1: lfp ~ 1
Model 2: lfp ~ k5 + k618 + age + wc + hc + lwg + inc
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       752    1029.75
2       745     905.27  7   124.48 < 2.2e-16 ***
```

small model: $\text{logit}(\pi) = \beta_0$ large model: $\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_7 X_7$

deviance 차이: $D_s - D_L = 124.48$ p값: $p < 2.2 \times 10^{-16}$

검정 결과: 귀무가설 기각 → 적어도 하나의 유의적인 모수 존재

- 참고

```
> summary(fit)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
k5          -1.462913   0.197001  -7.426 1.12e-13 ***
k618        -0.064571   0.068001  -0.950 0.342337
age         -0.062871   0.012783  -4.918 8.73e-07 ***
wcyes        0.807274   0.229980   3.510 0.000448 ***
hcyes        0.111734   0.206040   0.542 0.587618
lwg          0.604693   0.150818   4.009 6.09e-05 ***
inc         -0.034446   0.008208  -4.196 2.71e-05 ***
---
    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  905.27  on 745  degrees of freedom
AIC: 921.27
```

- Null deviance: 절편만 있는 모형, fit_0의 deviance
- Residual deviance: 현재 모형, fit의 deviance
- 직접계산으로 앞장과 동일한 검정실시 가능

2) $H_0: \beta_2 = \beta_5 = 0$ (예제 3.2 (2))

```
> fit_r <- glm(lfp ~ .-k618 -hc, family=binomial, data=Mroz)
> anova(fit_r,fit,test="Chisq")
Analysis of Deviance Table

Model 1: lfp ~ (k5 + k618 + age + wc + hc + lwg + inc) - k618 - hc
Model 2: lfp ~ k5 + k618 + age + wc + hc + lwg + inc
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       747     906.46
2       745     905.27  2   1.1895   0.5517
```

- p값이 0.5517으로 상당히 큰 값이므로 귀무가설을 기각할 수 없음
- k618과 hc는 모형의 적합도를 향상시키는데 도움이 안 되는 변수임

3) H_0: $\beta_1 = 0$

```
> fit <- glm(lfp ~ ., family=binomial, data=Mroz)
> fit_1 <- update(fit, . ~ . -k5)
> anova(fit_1, fit, test="Chisq")
Analysis of Deviance Table

Model 1: lfp ~ k618 + age + wc + hc + lwg + inc
Model 2: lfp ~ k5 + k618 + age + wc + hc + lwg + inc
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       746     971.75
2       745     905.27  1   66.484 3.527e-16 ***
```

Wald test와의 비교

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
k5          -1.462913   0.197001  -7.426 1.12e-13 ***
k618        -0.064571   0.068001  -0.950 0.342337
age         -0.062871   0.012783  -4.918 8.73e-07 ***
wcyes        0.807274   0.229980   3.510 0.000448 ***
hcyes        0.111734   0.206040   0.542 0.587618
lwg          0.604693   0.150818   4.009 6.09e-05 ***
inc         -0.034446   0.008208  -4.196 2.71e-05 ***
```

- 일반적으로 두 검정 방식의 결과는 동일함.
- deviance에 근거한 검정 방법이 더 안정적인 결과를 산출

- 예제 3.2 (1) $H_0: \beta_2 = \beta_5$

  - R에서 해결하는 방법을 아직 찾지 못했음
  - 가설 검정의 필요성 불분명

- 예제 3.3 (2) 설정된 모형이 완전정보를 가진 모형과 유의적인 차이가 있는지 데비언스 검정을 실시하라.

  - 이항반응변수의 경우 3.3절의 내용은 적용 불가
  - 예제 3.3(2)의 풀이 내용 무시하기 바람

2.4) Likelihood 함수에 기초한 개별 회귀계수의 신뢰구간 추정

- 신뢰구간 추정 방법

  1) Wald test에 기초한 방법: 직접 계산

$$\hat{\beta}_j \pm z_{1-\alpha/2} SE(\hat{\beta}_j)$$

  2) LRT에 기초한 방법
     - Profile likelihood에 의한 계산 방법
     - 이론적 배경 생략
     - R에서의 계산: 함수 confint( )

- LRT에 기초한 방법이 더 안정적인 결과를 산출

- 예제: 부인 직업 참여 자료

  - 회귀계수에 대한 95% 신뢰구간 추정

```
> confint(fit)
Waiting for profiling to be done...
                  2.5 %      97.5 %
(Intercept)  1.93697359  4.46630794
k5          -1.86089654 -1.08747196
k618        -0.19839650  0.06867096
age         -0.08830325 -0.03813509
wcyes        0.36099360  1.26377557
hcyes       -0.29200419  0.51679061
lwg          0.31402218  0.90697688
inc         -0.05099767 -0.01877093
```

  - 신뢰수준 조절:

```
> confint(fit, level=0.9)
                      5 %        95 %
(Intercept)  2.13496474  4.25679393
k5          -1.79524395 -1.14647112
k618        -0.17677870  0.04722581
age         -0.08415996 -0.04207230
wcyes        0.43221267  1.18953389
hcyes       -0.22704194  0.45143873
lwg          0.36019703  0.85738432
inc         -0.04827191 -0.02123953
```